Optimal algorithmic trading and market microstructure

Mauricio LABADIE

Charles-Albert LEHALLE

October 2010

Résumé

La frontière efficiente est un concept clé dans la Théorie Moderne du Portefeuille. En nous basant sur cette idée, nous allons construire des courbes de trading optimal pour des différents types de portefeuilles. Ces courbes correspondent aux stratégies de trading algorithmique qui minimisent l'espérance des coûts de transaction, i.e. l'effet joint de l'impact de marché et le risque de marché.

On va étudier cinq stratégies de portefeuille. Pour les trois premières (un seul actif, multi-actifs et portefeuille balancé) on assumera que les sous-jacents suivent une diffusion Gaussienne, tandis que pour les deux derniers on supposera qu'il existe une combinaison d'actifs telle que le portefeuille correspondant suit une dynamique de retour à la moyenne. Les courbes de trading optimal peuvent être calculées en resolvant un problème d'optimisation dans $\mathbb{R}^N$, où N est le nombre (pré-déterminé) de temps de trading. Dans quatre cas sur cinq, on obtient un simple algorithme récursif de la forme

$$x_{n+1} = F(x_n, x_{n-1}),$$

sous les contraintes $x_0 = 1$ et $x_{N+1} = 0$.

On va resoudre l'algorithme récursif en utilisant la *méthode de tir* (en anglais *shooting method*), une technique numérique des équations différentielles. Cette méthode a l'avantage que son équation correspondante est toujours unidimensionnelle, quoi qu'il soit le nombre de temps de trading N. De plus, cette technique peut être appliquée aussi à des portefeuilles plus généraux, pour lesquels l'équation a tant des dimensions comme le nombre de sous-jacents mais elle reste toujours indépendant de N.

Cette nouvelle approche pourrait intéresser des traders haute-fréquence et des courtiers électroniques.

Abstract

The efficient frontier is a core concept in Modern Portfolio Theory. Based on this idea, we will construct optimal trading curves for different types of portfolios. These curves correspond to the algorithmic trading strategies that minimize the expected transaction costs, i.e. the joint effect of market impact and market risk.

We will study five portfolio trading strategies. For the first three (single-asset, general multi-asseet and balanced portfolios) we will assume that the underlyings follow a Gaussian diffusion, whereas for the last two portfolios we will suppose that there exists a combination of assets such that the corresponding portfolio follows a mean-reverting dynamics. The optimal trading curves can be computed by solving an optimization problem in $\mathbb{R}^N$, where N is the (pre-determined) number of trading times. In four out of the five cases, we will obtain a simple, recursive algorithm of the form

$$x_{n+1} = F(x_n, x_{n-1}),$$

under the constraints $x_0 = 1$ and $x_{N+1} = 0$.

We will solve the recursive algorithm using the *shooting method*, a numerical technique for differential equations. This method has the advantage that its corresponding equation is always one-dimensional regardless of the number of trading times N. Moreover, this technique can be also applied to more general portfolios, for which the equation has as many dimensions as the number of assets but it is still independent of N.

This novel approach could be appealing for high-frequency traders and electronic brokers.

Contents

Chapter 1

Introduction

The Modern Portfolio Theory (MPT) and the Capital Asset Pricing Model (CAPM) are milestones in asset pricing and management in both the academy and the industry. These two theories are elegant theoretical achievements that have revolutionized the vision of Finance and Economics. We will review both theories in order to get some insight on the relationship between risk and return. Afterwards we will apply the same ideas for trading strategies in order to minimize the associated transaction costs.

1.1 Modern Portfolio Theory (MPT) and efficient frontier

MPT (or Markowitz Portfolio) was developed by Markowitz in 1952. The idea behind MPT is simple yet insightful. Imagine a market with two assets A and B, in which we invest today ($t = 0$) and at time $t = 1$ we recover our initial investment plus the profits of the period. Assume that the probability distributions of A and B are known, i.e. their means r_A, r_B and variances σ_A, σ_B are information available to everybody.

Suppose $r_A > r_B$ and $\sigma_A > \sigma_B$. Then we have two natural choices:

- Maximize profits regardless of the risk (i.e. variance). In this case we choose asset A.
- Minimize risk regardless of profit. In this case we choose B.

Now suppose that the correlation ρ between both assets is negative and that short-selling is not allowed. Then there exists an investment strategy $\omega \in (0, 1)$ such that the corresponding portfolio

$$P = \omega A + (1 - \omega)B$$

has minimal variance, i.e. $\sigma_P < \sigma_B$. Portfolio P is called the *minimal variance portfolio* (see Figure 1.1).

In general, if the market consists on N assets $A_1, \ldots, A_N$, there is an investment strategy

$$\omega_i \geq 0, \quad i = 1, \ldots, N; \quad \sum_{i=1}^{N} \omega_i = 1$$

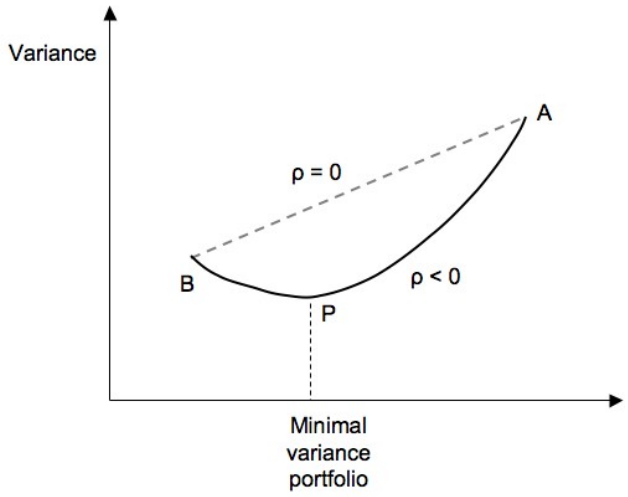

Figure 1.1: The effect of diversification. If the correlation $\rho = 0$ (dotted line) then the minimal variance portfolio is asset B. However, if $\rho < 0$ (solid line) there is a portfolio P with less variance than B (called minimal variance portfolio).

such that the portfolio

$$P = \sum_{i=1}^{N} \omega_i A_i$$

has minimal variance, i.e.

$$\sigma_P \leq \min\{\sigma_i \,:\, i = 1, \ldots, N\}. \tag{1.1}$$

Moreover, if at least one of the correlations is negative then inequality (1.1) is strict.

Now suppose we want to minimize the variance of our portfolio P for a given target return r. Then the optimization program is to minimize σ_P under the constraints

$$\omega_i \geq 0, \quad i = 1, \ldots, N; \quad \sum_{i=1}^{N} \omega_i = 1; \quad \sum_{i=1}^{N} \omega_i r_i = r.$$

Analogously, for a given risk level σ we can maximize the portfolio return r_p under the constraints

$$\omega_i \geq 0, \quad i = 1, \ldots, N; \quad \sum_{i=1}^{N} \omega_i = 1; \quad \sigma_P = \sigma.$$

Graphing the optimal pair (r_P, σ_P) we obtain a curve called *efficient frontier* (see Figure 1.2). Its name comes from the fact that the porfolios on it are the most efficient ones: they maximize returns for a given risk level, or equivalently, they minimize risks for a given target return.

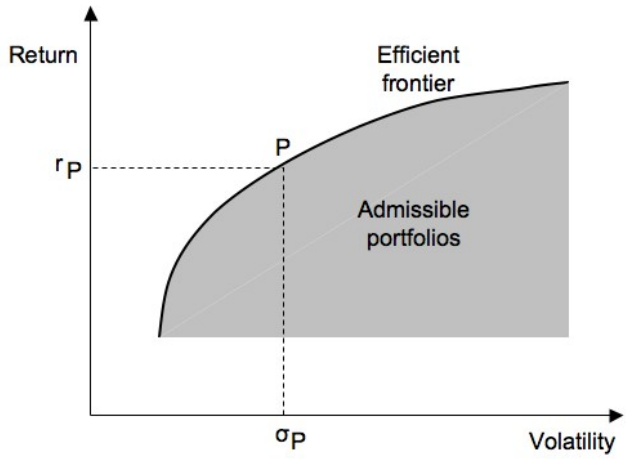

Figure 1.2: Efficient frontier. The curve separates the admissible portfolios (i.e. those satisfying the contraints) from the non-admissible ones. A portfolio P on the efficient frontier minimizes the risk (volatility) for a given level of return, or equivalently maximizes the return for a given level of risk.

1.2 Capital Asset Pricing Model (CAPM) and betas

MPT is a great idea that relies on the calculation of the variance-covariance matrix. However, when the number of assets grows it becomes very hard to calculate. Indeed, For N assets, since the $N \times N$ variance-covariance matrix is symmetric it has $N(N+1)/2$ degrees of freedom (See Table 1.1).

N	$N(N+1)/2$	Index with N components
5	15	-
10	55	-
15	120	-
20	210	-
30	465	Dow Jones, DAX
40	820	CAC 40
50	1,275	EUROSTOXX
100	5,050	FTSE
225	25,425	NIKKEI
500	125,250	S&P

Table 1.1: Even for the smallest indices (i.e. $N = 30$), the number of correlations that have to be calculated exceeds 450. This number quickly reaches 1000, even before $N = 50$.

In order to overcome this difficulty, we could try to calculate first a *market portfolio*, which includes all available assets, and then compare this market portfolio with each and every one

of the single assets. If we proceed this way then the number of degrees of freedom is $2(N+1)$: $N+1$ volatilities and $N+1$ correlations. This is far more manageable than the $N(N+1)/2$ degrees of freedom in MPT.

This is the idea behind CAPM, which was developed by Sharpe, a PhD student of Markowitz, in 1964. According to CAPM, the return of an asset i is

$$r^i = r^f + \beta_{iM}(r^M - r^f) + \varepsilon_i, \quad \beta_{iM} = \frac{\mathrm{cov}(r^i, r^M)}{\mathrm{var}(r^M)}, \tag{1.2}$$

where r^i is the return of asset i, r^f the return of the risk-free asset (e.g. Treasure bonds) and r^M the market return. β_{iM} is the marginal contribution of asset i to market risk, also known as the *systematic risk* or market risk, whereas ε_i is the *idiosyncratic risk*. The idiosincratic risk can be eliminated via diversification, whereas the systematic risk is inherent of the market and cannot be diversified away.

Now let us study the relative returns with respect to the risk-free asset. Taking expectations in (1.2) it follows that that the expected return of asset i over the risk-less rate r_f is

$$E(r^i - r^f) = \beta_{iM} E(r^M - r^f). \tag{1.3}$$

As we can see from (1.2), the beta of asset i (i.e. its systematic risk β_{iM}) acts as an amplifier of the expected market returns (see Figure 1.3).

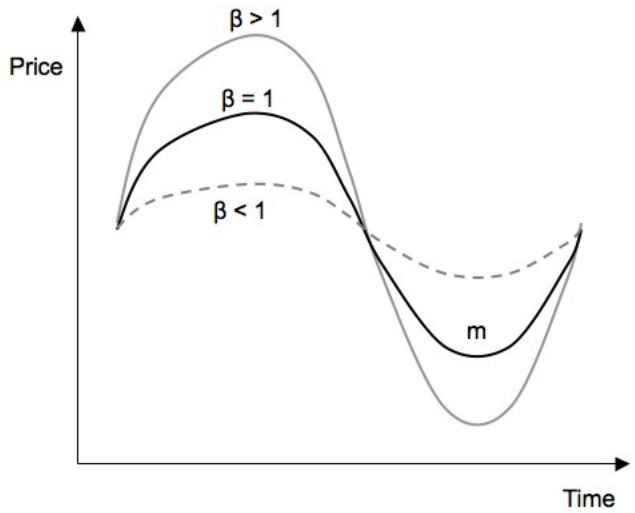

Figure 1.3: Beta. The market portfolio m has $\beta = 1$. For assets such that $\beta > 1$ both profits and losses are amplified, whereas for assets such that $\beta < 1$ both profits and losses are reduced.

1.3 Optimal trading curve

When it comes to intraday trading strategies we have the following dilemma, also known as the *trader's dilemma*: If we trade slow then prices will move away from their current quote, i.e. we are facing a *market risk*; however, if we trade fast then our order will drive quotes away from the current one, i.e. we will have a great *market impact* (see Figure 1.4).

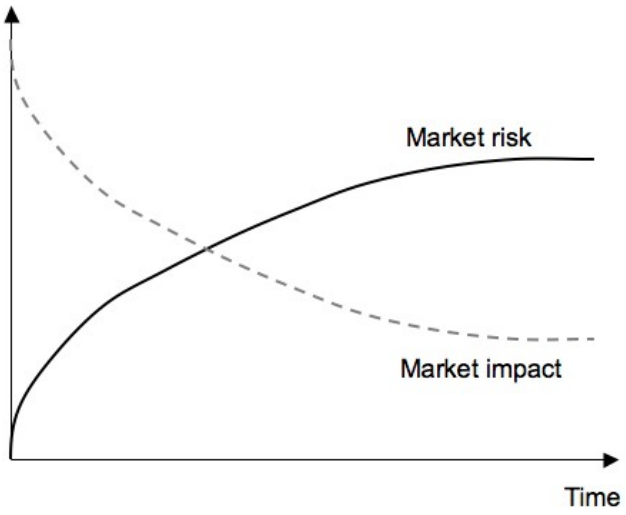

Figure 1.4: Trader's dilemma. Trading faster reduces market risk but increases market impact, whereas trading slower reduces market impact but increases market risk.

Recall that in MPT we optimize the joint effect of two oppossite forces: minimizing the risk of the portfolio and maximizing the (expected) return. Following the idea of the efficient frontier, it seems natural to build up an optimization program that minimizes simultaneously both the market risk and the market impact.

Suppose we need to sell a certain amount of asset S during the day. We split the trading order in exactly N small sub-orders of size ν_n, $n = 1, \ldots, N$. The goal is to find the right trading proportions

$$\nu_i \geq 0, \quad i = 1, \ldots, N; \quad \sum_{n=1}^{N} \nu_n = 1,$$

that minimize the expected loss due to market risk and market impact.

As we will see in later chapters, the set of minimizers constitute a curve, the *optimal trading curve*. For a given risk level (variance), the trading strategy P on the optimal trading curve is the one that minimizes the expected market costs, i.e. the joint effects of market risk and market impact. Conversely, given a level of expected market costs, the optimal strategy P minimizes the market risk (variance) (see Figure 1.3).

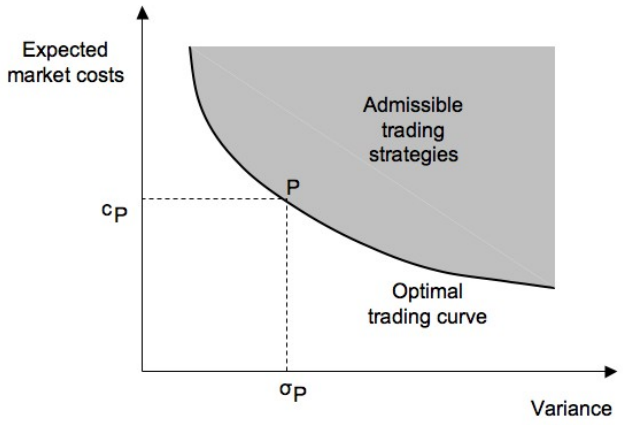

Figure 1.5: Optimal Trading curve. Trading strategies P on the curve minimize the joint effect of market risk (variance) and market impact (expected market costs).

The optimal trading strategy is thus the vector of proportions $(\nu_1, \ldots, \nu_N)$ that must be exchanged at each trading time. It is customary to describe trading curves not in terms of the number of assets exchanged but in terms of the remaining assets in the portfolio:

$$(x_0, \ldots, x_{N+1}), \quad x_0 = 1, \quad x_{N+1} = 0, \quad x_n = \sum_{i=n}^{N} \nu_i \quad \forall n = 1, \ldots, N.$$

1.4 The scope of this mémoire

The goal of this mémoire is to describe thoroughly the construction of the optimal trading curve $(x_0, \ldots, x_{N+1})$ for different market models and portfolio strategies.

In Chapter 2 we will study the market microstructure. We will see how the hypotheses of MPT and CAPM, i.e. the Efficient Market Theory, are all violated in real markets. We will focus in particular on the effect of transaction costs and market impact. We will also review the benchmarks used for monitoring trades.

Roughly speaking, a trading strategy is *algorithmic* if it is stripped of human decisions (and emotions). In Chapter 3 we will describe what is algorithmic trading and we will survey the basic strategies in algorithmic trading, which are the building blocks of almost any systematic trading strategy can be constructed. We will also show evidence that favors algorithmic trading over human trading.

In Chapter 4 we will construct the optimal trading curve $(x_0, \ldots, x_{N+1})$ under normality assumptions, i.e. the asset is supposed to follow a Brownian motion. This chapter will be based on the article of Almgren and Chriss [1] for single assets and on the work of Lehalle [14] for

multi-asset and balanced portfolios.

In Chapter 5 we will construct again the optimal trading curve $(x_0, \ldots, x_{N+1})$, but following Lehalle [14] we will consider that the portfolio has a mean-reverting dynamics. We will solve analytically and numerical a simplified case of a mean-reverting portfolio using the *shooting method*, a numerical technique used in differential equations. The novelty of our approach is the alternative optimization program we use: we will construct the optimal trading curve using a 1-dimensional algorithm regardless of the total number of trades N. Being more advantageous than the classical approaches based on functional optimization in $\mathbb{R}^N$, this approach could be of interest for systematic brokers and traders.

Chapter 6 is the final chapter. We will make some remarks on the portfolio models we have presented and mention some possible extensions. We will also review several alternative models for time series that could be used to describe markets more accurately. Finally, we will comment on the pros and cons of automated (algorithmic-based) trading with respect to discretionary (human-based) trading.

Chapter 2

Market microstructure

2.1 Hypotheses behind MPT and CAPM: The Efficient Market Theory

Despite the beauty and simplicity of MPT and CAPM, the theory they rely on, i.e. the Efficient Market Theory (EMT) is too reductionistic and idealistic when compared with real market conditions. Therefore, MPT and CAPM must be handled with care since they both can lead to wrong conclusions.

Let us study each one of the hypotheses of the EMT, the framework in which MPT and CAPM were developed.

1. *Existence of a single market price.*

   According to the theory, market prices reflect the *fundamental value* of assets. However, the very notion of price is very ambiguous. Indeed, in any market we have several prices coexisting simultaneously: ask price, bid price, mid-point, last traded price, average price, etc. Moreover, this single-price assumption ignores the price formation process, which depends on the subtleties of each market and explains why do we have different prices at different markets and .

2. *Information is complete and perfect.*

   According to EMT, economic information is complete, perfect and everyone has access to it. Therefore, if investors are rational they will all have the same expectations on the future behavior of assets. In practice this is not true because there exists an asymmetry of information. Indeed, not only information has a price (e.g. real-time access via Bloomberg or Reuters) but also markets have different degrees of transparency (e.g. dark pools).

3. *All investors are equal.*

   If all investors were rational and share the same information then they would all have the same expectations on the future value of assets, and in consequence they would all have the same behavior. However, since there is a huge heterogeneity of investors it is not

realistic at all to consider that all investors are equal, as the EMT does. Indeed, each single investor has a personal strategy (long-only, long-short, hedging, speculation, arbitrage), a time horizon (ranging from several years to milliseconds) and an asset preference (equities, foreign exchange, interest rates, credit, derivatives, venture capital).

4. *Agents are infinitely rational.*

   All agents (i.e. market participants) are suppose to have a utility function that describes all their preferences, which they try to maximize. This hypothesis rises two questions. On the one hand, investors do have personal biases due to their beliefs (politics, culture and religion), which are hard to quantify. On the other hand, there is abundant evidence of herd behavior and self-fulfilling anticipations.

5. *No endogenous crashes.*

   The EMT affirms that market prices reflect the *fundamental value* of the assets, and that these prices only move due to unpredictable events or news. Under this framework, crashes can only be exogenous, never provoked by the inner dynamics of the markets. However, in the past hundred years we have had several crashes, most of them caused by the markets themselves : the Great Depression in 1929, the "Black Monday" on October 19 1987, the Internet bubble in 2000, the subprime in 2008 and the flash crash on May 6 2010.

Since none of these hypothesis is fully verified in real markets, it is important to be aware of the limits of the EMT approach. This is particularly true for constructing market models, especially if the goal is to exploit trading opportunities. The discipline known as *market microstructure* aims to understand the effect of these factors (among others) in order to better understand the markets.

2.2 Market structure

Among the microstructure effects, market structure is one of the most important ones, just behind transaction costs. Unlike the assumptions of the EMT, where all markets are treated in a democratic fashion, the microstructure theory states that the specific organization of each market determines the price-formation processes and its intrinsic trading dynamics.

Understanding the way each market works is crucial for all traders, especially high-frequency traders who try to reap profits from small anomalies in intraday prices, without being exposed to market trends. Here we will survey the different kinds of markets and orders. For further references we invite the reader to check Barry Johnson [13] and Fabrice Riva [21].

2.2.1 Market types

There are three types of markets: order-driven markets, quote-driven markets and hybrid markets.

In an **Order-driven market** all traders participate equally, placing orders on an order book that are matched following a consistent set of priority rules. In general, first and second priority in the order book are given to price and time, respectively. However, there are markets with a membership priority, which is placed between price and time. A remarkable feature of order-driven markets is their degree of automatization, which makes this kind of markets very appealing for automated trading, in particular for high-frequency strategies.

In a **Quote-driven market** traders must transact with dealers or market makers who quote prices at which they will buy and sell a given quantity. Since the role of the market maker is to provide liquidity, the prices they quote are firm.

A **Hybrid market** is in principle an order-driven market, but they allow direct negotiation between counterparties if the trading volumes are sufficiently big.

2.2.2 Tick and fixing

Independently of their kind, all markets share two specific intraday trading characteristics, the tick and the fixing.

A **tick** is the smallest price change available. As a rule of thumb, the more liquid the asset is, the tighter its bid-ask spread is. However, if the tick size is too big then even for the most liquid assets the spread will be large because it is always a multiple of the tick. The tick size also affects the volatility: if the tick is big then even the slightest change in price has a strong impact.

A **fixing** is a market period with discontinuous price quotes. It has two stages: a *pre-fixing* period when buy/sell orders are cumulated but not executed, and a *trading* period where all cumulated orders are traded at a fixed price. In practice, this unique price is such that the number of exchanges is maximized, i.e. it is a walrasian equilibrium price. There are several reasons for a fixing: during the opening, it allows a more efficient price discovery mechanism, whereas for the closing it reduces the volatility and the price manipulation. There also exist non-schedule fixings. For example, if the volatility is too high then the stock exchange can call a fixing, which lasts 5 minutes. This permits to stop the abnormal price trend and facilitates the return to the normal equilibrium.

2.2.3 Market orders

A **Market order** is an instruction to trade a given quantity at the best price possible. Market orders demand liquidity because their focus is on completing the order. Therefore, the main risk is the uncertainty of the ultimate execution price: if the volume at the current market price is not enough then the market order jumps to the next level of the order book; this process goes on until the order is fully executed.

A **Limit order** is an instruction to trade a given quantity at a specified price or better. A buy limit order must execute at or below this limit price, whereas a sell order must execute at

or above its limit price.

A **Market-to-limit order** is an hybrid instruction constituted by a market order with an implicit price limit. When the order arrives it behaves as a market order, seeking liquidity at the best price available, which we call the *entry price*. As soon as the order starts to execute, it becomes a limit-order with limit price equals to the entry price. Unlike a traditional market order, a market-to-limit order does not sweep the order book. If there is insufficient liquidity available at the best price, the order will convert into a standing limit order for the residual amount.

A **Stop order** is an extension of the market-to-limit order with a limit price further away from the last execution price, i.e. the trading is activated or stopped when a certain threshold price is reached. There are three important examples of stop orders. **Stop-loss** orders are designed to protect a potential gain: for buy (resp. sell) orders the execution is stopped if prices go above (resp. below) the threshold. **Contingent or if-touched orders** remain hidden until the threshold is reached, in which case they become active; hence, they are the mirror orders of stop-loss. **Stop limit** orders have two thresholds, one that activates the order and the other one that deactivates it, and as such they are a hybrid built with one stop-loss and one contingent order.

An **Iceberg order** is an order with a small part visible in the order book and a significantly larger hidden volume. These orders slice the total amount to be exchanged into several tranches. The first tranche constitutes the visible part, and as soon as it is completely executed the next tranche becomes visible. The interest of iceberg orders is that they provide an automated slicing program for orders of big size. However, the hidden tranches lose time priority in the order book; they only have price priority.

A **Peg order** is an instruction with a dynamic limit price. The price is automatically adjusted according to the evolution the spread: for buy (resp. sell) orders they always hit the best bid (resp. ask) price. In consequence, peg orders are always first in price priority and second in time priority. There are also peg orders with stop-limits, which follow the best price until it reaches the deactivating threshold.

2.3 Transaction costs

Transaction costs are by far the most important of the microstructure effects, not only because they are determinant factors in high-frequency trading but als because all markets particularities and trading mechanisms can be interpreted as friction factors, and as such they are included in the transaction costs. A profitable trading strategy in theory could not be so in practice because transaction costs are bigger than the expected gains. Therefore, it is important to quantify the transaction costs before launching the trading strategy, in particular for the most active ones.

Transaction costs can be explicit or implicit, but in any case it is important to monitor them

and have an idea of the impact they have on the trading strategy. We will survey the most important transaction costs from the most visible ones to the most hidden ones.

Commissions, which are the most obvious cost component, represent the broker compensation.

Fees are charges that may be levied by floor brokers and exchanges. They also include the costs of clearing and settlement. In general, fees are already included in the comission charge.

Taxes are generally charged on the realized profits from capital gains. However, there are stock markets (e.g. UK) where there is a duty on share purchases for non-members, and there is a current debate on whether this measure should be extended to other markets.

Spread costs are the only compensation that market makers and traders receive for providing liquidity. As we have mentioned before, the spread usually reflects the liquidity of an asset and strongly depends on the tick size. For single executions, spreads are straightforward to calculate, but for trades that are split up into several small orders we need to track down the spread costs for each executed order separately.

Delay costs reflect any price change between the time when the decision to invest was made and the actual time when the order started to be executed. Delay costs can be a substantial proportion of overall costs, particularly for assets with high volatility and/or whose price is trending unfavorably.

Market impact represents how much effect does the trading order have on the price: larger orders will result in a bigger impact than smaller ones. This effect decreases significantly as liquidity (i.e. the available trading volume) increases. Market impact can be split into temporary and permanent. The temporary market impact reflects the overall cost incurred by demanding immediacy. The permanent impact carries the information leakage: the trade unveils a different long-term price expectation, hence the price is adjusted accordingly in the long term.

Price trends can affect a trade in a significant way. Indeed, if there is an upward trend then the best strategy is to buy early and sell later, whereas for a downward trend we should sell early and buy later. Therefore, if we entered into the wrong position we will be paying the trend, especially if the benchmark we are using to monitor our trade is the closing price.

Market risk or timing risk reflects the uncertainty of prices as time passes. It is related to volatility because the more volatile is an asset, the more likely it is to drift away from its initial price. Therefore, market risk can have a considerable cost for strategies based on volatile assets and/or orders that have a long trading horizon.

Opportunity costs represent missed opportunities due to incomplete trading orders or unfavorable market conditions. This costs are generally tracked using a *paper portfolio*, i.e. a theoretical, ideal trading strategy that we are trying to replicate.

In the next chapters we will build several microstructure models focussed on market impact and market risk. It is important to notice that we are not losing any generality because fixed and liquidity costs (commisions, fees, taxes and spreads) can be incorporated in the market impact function, whereas delay and opportunity costs can be related to market risk.

2.4 Monitoring trading: Benchmarks

Choosing the right benchmark is of utmost importance because it is the measure stick that determines whether a trading strategy is profitable or not. A good benchmark should be easy to track, verifiable and provide an accurate measurement. There are three kinds of benchmarks: pre-trade, intraday and post-trade.

2.4.1 Pre-trade benchmarks

They have the advantage that they are easily determined and immediately available for comparison. The most used ones are **previous close** (i.e. last quote yesterday), **opening price** (first quote today), **decision price** (price at the moment when the investor decided to trade) and **arrival price** (price when the broker started to execute the order).

Since pre-trade benchmarks do not influence or get influenced by the market, they are objective measures for transaction costs. However, a substantial price shift during the day can make the bechmark less meaningful because the traded orders will be compared to a price they could never have achieved.

2.4.2 Intraday benchmarks

They are average prices that try to reflect more accurately the intraday market conditions than pre-trade benchmarks do.

OHLC is the average of four numbers: Open, High, Low, and Close. It used to be a proxy for mean market price, but given the available amount of data nowadays it has lost its appeal. Notice that OHLC being an average of only four points, it is only meaningful if the market is not volatile because it can be easily distorted by extreme values.

TWAP (Time Weighted Average price) is an average of the observed trade prices over a time period. TWAP is a dynamic benchmark because for each new trade it incorporates a new price to the existing ones and updates the average. Notice that since all values for TWAP have the same weight, extreme prices can have a large effect on it.

VWAP (Volume Weighted Average Price) is the total traded value over the total traded quantity, and as such it gives the fairest indication on how do market prices have moved over a time period. Unlike TWAP, where all prices have the same weight, VWAP weights each traded price by its corresponding traded size. Therefore, small trades at extreme prices are smoothed

out whereas the largest trades will dominate the average. However, VWAP is not very accurate for large trades. Indeed, if an order represents an important percentage of the day's trading (i.e. over 30%) then VWAP has no meaning as a performance measure because the trade will have a great impact on the average. VWAP can also lead to artificially poorer performances for assets that are volatile or markets that have a strong trend.

2.4.3 Post-trade benchmarks

We have two important post-trade benchmarks: closing and implementation shortfall.

Closing prices are a very popular benchmark, widely used as a milestone for marking to market and for profit and loss because it is a succint summary of price changes during the day. However, closing prices have the same disadvantage than opening prices: they do not reflect the trading conditions trhoughout the day. In consequence, although closing prices are a popular benchmark, they are not the best reference for performance analysis.

Implementation Shortfall (IS) is a measure of the total transaction costs. It consists on comparing the actual performance of the portfolio with its *paper* equivalent, i.e. a virtual portfolio traded at benchmark prices:

$$IS = \text{Returns of paper portfolio} - \text{Returns of real portfolio.}$$

In order to describe IS more precisely, let us first make some definitions. Let X be the intended investment, p_D the price when the investment decision was made, p_F the final market price, p_A the price when the order started to be executed (i.e. the arrival price), ν_n the size of the individual executions and p_n the achieved prices. Under this framework, the transaction costs are

$$IS = \underbrace{X(p_A - p_D)}_{\text{Delay costs}} + \underbrace{\sum_{n=1}^{N} \nu_n (p_n - p_A)}_{\text{trading costs}} + \underbrace{\left(X - \sum_{n=1}^{N} \nu_n \right) (p_F - p_A)}_{\text{opportunity costs}} + \text{explicit costs.}$$

Observe that if all the order was executed, i.e.

$$X = \sum_{n=1}^{N} \nu_n,$$

then the only transaction costs are the delay and explicit costs,

$$IS = \sum_{n=1}^{N} \nu_n (p_n - p_D) + \text{ explicit costs.}$$

Chapter 3

Algorithmic trading

3.1 Some facts on algorithmic trading

There are two different classifications of trading strategies: on the one hand we have algorithmic vs human trading, whereas on the other hand we have high vs low frequency trading. It is important to stress that both classifications are not at all exclusive. Indeed, a high frequency trader can be algorithmic or human as long as it trades actively; on the contrary, an algorithm can decide to trade and hold a position on an illiquid asset for a long time horizon.

There is no strict distinction between high and low frequency trading. However, as a rule of thumb , practitioners set the threshold at 15 minutes: we are in the high frequency regime if the time between trades is less than 15 minutes and in the low frequency regime if the time between trades is greater than 15 minutes (see Lehalle and Rosenbaum [15]).

3.1.1 Empirical evidence favoring algorithmic trading over human trading

According to Hendershott and Riordan [12], high-frequency trading (HFT) has currently stronger presence than any other trading strategy (discretionary trading included). The authors present two main reasons to support this statement. First, HFT firms are just 2% of the 20,000 trading firms in the US but nevertheless they hold 73% of stock trading in the US and 40% in Europe. Second, in the German stock market (Deutsche Boerse), automated HFT dominates Human HFT: algorithms count for more than 52% of total liquidity demand and 50% of total liquidity offer.

Hendershott and Riordan also present empirical evidence that algorithmic traders are more efficient than human traders. They monitored the German stock market and compared the behavior of algorithmic and human traders. On Xetra, the electronic trading platform of the German stock market, they observed that algorithms and humans have different trading patterns:

- Algorithmic traders are more present in transactions of smaller size than humans and their

participation rate is decreasing in size.

- Human traders are more present in transactions of bigger size than algorithms and their participation rate is increasing in size.
- Given an algorithmic order of very small size, there is a 48% probability that the trade that follows will be very small and algorithm-generated as well.

Concerning the behavior of algorithmic traders in terms of liquidity and price discovery, the authors found the following facts:

- Algorithmic traders improve the bid-ask spread: they are liquidity consumers when it is cheap (i.e. narrow spreads) and providers when it is expensive (i.e. big spreads). Moreover, they are within the spread more frequently than humans.
- Algorithmic traders contribute for 51% to price discovery whereas human contribute for only 39%.
- Automated traders do not increase market volatility. All the opposite, they lower the volatility by keeping steady the available liquidity level.

3.2 Algorithmic trading and its multiple faces

Let us describe the different types of algorithmic trading.

Systematic trading, also called *automated trading*, consists on adopting the same approach for each trade, i.e. following a pre-determined set of rules to trade in a specific way. For example, we can fix entry or exit thresholds in order to start or stop trading. The rules behind a systematic trading algorithm can be simple or complex, but the nature is always the same : to repeat a given strategy over and over again.

In **Quantitative trading**, sometimes referred to as *black-box trading*, the trading rules are enforced by adopting propietary quantitative models. Such models are called *black-boxes* since they are closely guarded and only known by a few. Quantitative trading is sometimes confused with algorithmic trading. However, the former instigate trades whereas the latter merely execute them. In consequence, the goals of quantitative trading are wider than those of systematic trading. Quantitative trading is considered the most scientific kind of trading because traders build their own market models in order to define trading strategies. In consequence, quantitative trading aims to make money because of a better understanding of markets, whereas discretionary trading prefers a heuristic approach based on experience.

High frequency trading aims to take advantage of opportunities intraday. This kind of trading is a specialized form of quantitative trading focussed on exploiting short term gains. Frequently, high frequency traders take relatively market neutral positions, i.e. a net exposure to market equal to zero, whose only goal is to profit from price discrepancies (i.e. market microstructure). It is important to stress that high frequency traders need instant response to

market changes; in fact, the decision-making time in HFT is measured in milliseconds. This is the main reason why they invest on huge computing power and on direct market access (e.g. they rent clusters next to the stock exchanges they trade).

Statistical arbitrage represents a systematic trading approach based on a fusion between real-time and historical data analysis. The goal is to take advantage of the mispricing of financial instruments while minimizing overall risk. Frequently, statistical arbitrage relies on black-box strategies based on innovative tools borrowed from Economics and Science, e.g. time series, data mining, artificial intelligence, agent-based models and fractals.

3.3 Basic bricks for algorithmic trading

There are hundreds of different trading algorithms available. However, we can decompose them into a handful of basic algorithms regrouped in three main categories: impact-driven, cost-driven and opportunistic.

3.3.1 Impact-driven algorithms

They aim to minimize market impact by slicing a big trading order into smaller *child* orders. The core impact-driven algorithms here are three: TWAP, VWAP and POV.

Time Weighted Average Price (TWAP) slices one big order of size X into N equal child orders of size X/N, which are uniformly traded during a pre-determined time horizon. This means that each time-step a child order of size X/N is executed, independently of both market and volume price. Since this trading algorithm is extremely predictable, common variation of TWAP include small random perturbations of both the pre-determined trading times and child order size to reduce detectability.

Volume Weighted Average Price (VWAP) corresponds to the overall turnover divided by the total volume, hence large trades have more impact on the benchmark price than small ones. Unlike TWAP, when the only thing that matters is to trade regularly throughout the day, for VWAP we also need to trade in the right proportions, which are defined statistically following historical volume profiles. The performance of VWAP algorithms depends on how historical volume profiles are calculated and how well they forecast market volumes. In consequence, they are vulnerable to sudden shifts in trading volume or liquidity.

Percent of Volume (POV) aims to trade a fixed percentage of the current market volume. Unlike TWAP or VWAP, where the trading pattern is predetermined, the trading schedule for POV is dynamically determined. Since POV does not predict market volumes, common variations include historical analysis of volume profiles in order to anticipate upcoming trading volumes.

3.3.2 Cost-driven algorithms

They try to reduce the effect of overall transaction costs such as market impact and market risk.

Implementation Shortfall (IS) represents the difference between the price at which the investor decides to trade and the average execution price that is actually achieved, hence IS algorithms strike the right balance between market impact and market risk. Frequently, IS algorithms trade most of the order at the beginning of the execution because prices are closest to their benchmark. There are three optimization steps to do for IS algorithms. First, we determine the optimal trading horizon, depending on the order size and historical market volumes. Second, we decide the trading schedule, e.g. the number of trades and the time between trades. Third, we determine the sizes of each one of the child trades.

The trading strategies of the following chapters are all IS algorithms. We will assume that both the optimal trading horizon and schedule are already determined, so we will only have to worry about the optimal slicing of the order.

Market Close (MC) is an algorithm used when the benchmark is the close price, as it is the case of many trading firms. Since the bechmark is unknown until the end of the day, MC algorithms will try to trade near the closing time. MC is the mirror of IS: just as IS algorithms determine the optimal ending time, MC algorithms calculate the optimal starting time. Common variations of MC include a minimum (or maximum) order size allowed to participate in the close auction.

3.3.3 Opportunistic algorithms

They seek to take advantage whenever market conditions (price, liquidity, volatility or another factor) are favorable.

Price Inline (PI) algorithms are based on an impact-driven trading algorithm (e.g. VWAP or POV) to which they add a price adaptive functionality (i.e. a price sensitivity factor). PI algorithms adapt their strategy to the market price in a similar way to how POV algorithms adjust to market volume: they alter their trading pattern based on how does the market price compare to the benchmark. Common variations include aggressive or passive behaviors, depending on whether the participation rate is increased when prices are favorable or unfavorable, respectively. An aggressive strategy assumes that trends are short-lived and will soon revert whereas passive strategies rely on the trend persisting. Therefore, the choice between aggressive and passive relies on an a priori on the market behavior.

Liquidity-driven (LD) algorithms are the second-generation of simple rule-based order routing. The current world market is fragmented because there are several execution venues that compete with the traditional stock exchanges. In consequence, liquidity has become a crucial asset, and as such it must be tracked closely. LD algorithms aim to capture the liquidity as soon as it appears. In order to do so, LD algorithms follow in general two rules: First, they slice

the original order into small child orders that are sent to different execution venues. Second, they search hidden liquidity on each of the venues (e.g. iceberg or pegged orders). If the trading constraints imposed on a LD algorithm are too tight, there is a risk that the full order will not be completely executed. In practice, LD algorithms have some finish-up logic that avoids this scenario. It is worth to mention that the original LD algorithms were designed for trading illiquid assets, for which the signalling (i.e. informational) risk is a determinant factor.

Pair trading (PT), by far the basic tool for statistical arbitrage, consists on buying one asset while simultaneously selling another. This strategy is market neutral because the risk on the long side offsets the risk on the short side. However, this strategy only works if the two chosen assets are sufficiently correlated. PT assumes that the spread (i.e. the difference in prices) has a mean-reverting behavior. When the spread crosses a threshold, a trading signal is activated: if the spread goes above the upper threshold we sell the spread, whereas if it goes below the lower threshold we buy the spread. We exit the position when the spread crosses the benchmark (see Figure 3.1). Traders use different statistical tools to determine the benckmark and the upper and lower thresholds. The simplest case is to choose the mean (or a moving average) as benchmark, whereas the upper and lower thresholds are placed two standard deviations above or below the man, respectively.

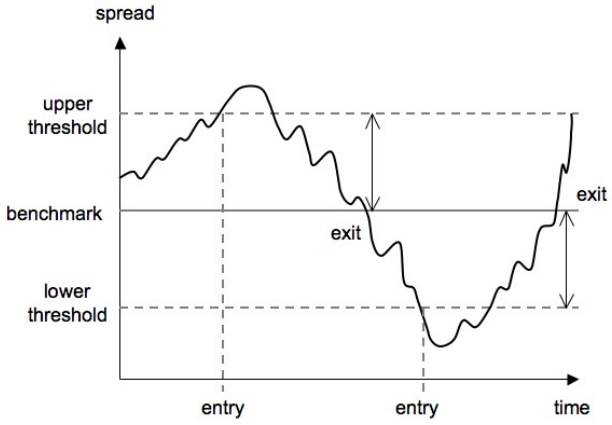

Figure 3.1: Pair trading. When the spread leaves the stripe between the upper and lower threshold we enter the position, i.e. we trade the spread. We exit the position as soon as the spread coincides with its mean, making a profit equal to half the stripe (double-headed arrows). This process is repeated throughout the day.

3.4 Building complex algorithms

Using the basic algorithms we have just described we can construct more complex algorithms. We will mention here three of the most common ones.

Portfolio trading, also known as basket or program trading, provides investors with the alternative to trade multiple assets in one go rather than trading them individually. This type of trading is provided by brokers, who generally use automated trading in order to execute the orders. There are two main advantages for portfolio trading over trading single stocks. First, trading a whole basket of assets is more cost-effective than trading the assets individually. Second, a portfolio has a diversification effect that tends to decrease its volatility.

In the following chapters we study in more detail some of these portfolio trading strategies. They belong to the IS category because they aim to minimize the joint effect of market risk and market impact.

Multi-leg (ML) is the next level of complexity of pair trading. It consists on trading several assets according to a benchmark price: we go long the undervalued assets and short the overvalued ones. They are widely used for *term-structure arbitrage*, i.e. we trade the same assets but with different maturities (e.g. bonds, futures and options). ML algorithms are also used for *carry trade* (borrow at a low interest rate and deposit on a high one) and for trading specific optional profiles (e.g. bull/bear spreads, collars, butterflies, condors, straddles and strangles).

Multi-asset (MA) algorithms started with derivatives trading. Derivatives depend on several factors such as underlying price, volatility, dividends, interest rates and time to maturity. Sometimes a trader wants to *hedge* the derivative, i.e. neutralize its sensitivity with respect to one of the factors. For example, *Delta-hedging* consists on trading a certain amount of the underlying, which in the right proportions eliminates the sensibility of the option to changes on the underlying. There is also Vega-hedging (with respect to volatility), Rho-hedging (time to maturity) and Gamma-hedging (big changes in the underlying), for which it is necessary to include other assets to the portfolio.

Investment banks used MA trading to hedge their positions, not only in vanilla derivatives but also in structured products, which are a combination of cash and derivatives. Two examples of these structured products are principal protected notes (bond + option) and equity linked notes (corporate bond + option + credit default risk). Hedge funds and propietary trading desks are also interested in MA trading for cross-asset arbitrage. Some of the most used strategies are index arbitrage (futures vs stocks), basis trading (futures vs bonds), option arbitrage (options vs stocks) and credit arbitrage (CDS or convertible bonds vs stocks).

Chapter 4

Optimal trading for Gaussian assets and portfolios

As we have stressed in previous chapters, the performance of trading algorithms depends on how well they process historical and real-time data and predict future market conditions. It is therefore evident that the accuracy of the forecasting model is of paramount importance.

As we mentioned before, when we trade we try to do so in an optimal way, i.e. minimizing simultaneously our market impact and market risk. To achieve this goal under the framework of systematic trading strategies, we need to construct an algorithm with market data as inputs and trading sizes as outputs. In this chapter we will construct two models for Implementation Shortfall algorithms, for which we will assume that the assets follow a Gaussian diffusion.

4.1 Single assets

4.1.1 Model

We will suppose that all assets are Gaussian. More precisely, for any asset S_n where $n = 1, \dots, N$ are the trading times, we have

$$S_{n+1} = S_n + \sigma_{n+1}\varepsilon_{n+1}, \tag{4.1}$$

where $\sigma_n > 0$ for all n and $(\varepsilon_n)_{1\leq n\leq N}$ are i.i.d.[1] Normal random variables of mean zero and variance 1. Following Almgren and Chriss [1], we will model the market impact as a function h depending on the average trading on each interval. More precisely, let

$$(\nu_1, \dots, \nu_N), \qquad \sum_{n=1}^{N} \nu_n = 1$$

be the number of units we trade at each time n. Then our assumption is that, for any n, $h_n = h(\nu_n)$.

[1] Independent and identically distributed.

Under this framework, we can calculate the wealth process (i.e. the full trading revenue upon completion of all trades):

$$W = \sum_{n=1}^{N} \delta_n \nu_n (S_n + \delta_n h(\nu_n)), \tag{4.2}$$

where $\delta_n = 1$ if we buy at time n and $\delta_n = -1$ if we sell. To simplify the analysis we will suppose that the trading strategy is long-only, i.e. $\delta_n = 1$ for all n. We will also assume that the market impact function h is linear and piecewise constant. More precisely, we will suppose that

$$h(\nu_n) = \eta \frac{\sigma_n \nu_n}{V_n}, \qquad \eta > 0, \tag{4.3}$$

with V_n the historical volume of the asset at time n. It is worth to remark that in Almgren and Chriss [1] the market impact is $h(\nu_n) = \eta \nu_n$. However, in (4.3) we are assuming that the market impact is proportional to the volatility of the asset and inversely proportional to the available volume, as it is done in Lehalle [14].

Under these conditions, the wealth process (4.2) in its simplified form is

$$W = \sum_{n=1}^{N} \nu_n S_n + \eta \sum_{n=1}^{N} \frac{\sigma_n}{V_n} \nu_n^2. \tag{4.4}$$

In order to optimize (4.4) it is desirable to eliminate the dependence of W with respect to the random process $(S_n)_{1 \leq n \leq N}$.

Using

$$S_n = S_0 + \sum_{i=1}^{n} \sigma_i \varepsilon_i$$

we have

$$\sum_{n=1}^{N} \nu_n S_n = \sum_{n=1}^{N} \nu_n S_0 + \sum_{n=1}^{N} \sum_{i=1}^{n} \nu_n \sigma_i \varepsilon_i.$$

However, a simple computation shows that

$$\sum_{n=1}^{N} \sum_{i=1}^{n} \nu_n \sigma_i \varepsilon_i = \sum_{n=1}^{N} \sigma_n \varepsilon_n \left(\sum_{i=n}^{N} \nu_i \right).$$

Therefore, if we define x_n as the remaining quantity of asset to buy at time n, i.e.

$$x_n = \sum_{i=n}^{N} \nu_i, \quad \nu_n = x_n - x_{n+1},$$

we obtain that

$$\sum_{n=1}^{N} \nu_n S_n = S_0 + \sum_{n=1}^{N} \sigma_n \varepsilon_n x_n.$$

In summary, (4.4) can be rewritten as a function of $(x_1, \ldots, x_n)$, i.e.

$$W(x_1, \ldots, x_n) = S_0 + \sum_{n=1}^{N} \sigma_n \varepsilon_n x_n + \eta \sum_{n=1}^{N} \frac{\sigma_n}{V_n}(x_n - x_{n+1})^2. \tag{4.5}$$

4.1.2 Optimization program

We compute the expectation and variance of the wealth process (4.5) as functions of x_n,

$$\begin{aligned} \mathbb{E}(W) &= S_0 + \eta \sum_{n=1}^{N} \frac{\sigma_n}{V_n}(x_n - x_{n+1})^2, \\ \mathbb{V}(W) &= \sum_{n=1}^{N} x_n^2 \sigma_n^2. \end{aligned} \tag{4.6}$$

Our goal is to find the trading strategy $(x_1, \ldots, x_N)$ that minimizes the expected market costs, represented by the wealth process W, for a given level of risk aversion λ. Therefore, we have to minimize the cost functional

$$\begin{aligned} J_\lambda(x_1, \ldots, x_N) &= \mathbb{E}(W) + \lambda \mathbb{V}(W) \\ &= S_0 + \eta \sum_{n=1}^{N} \frac{\sigma_n}{V_n}(x_n - x_{n+1})^2 + \lambda \sum_{n=1}^{N} x_n^2 \sigma_n^2. \end{aligned}$$

The partial derivative of J_λ with respect to x_n is

$$\frac{\partial J_\lambda}{\partial x_n} = -2\eta \frac{\sigma_{n-1}}{V_{n-1}}(x_{n-1} - x_n) + 2\eta \frac{\sigma_n}{V_n}(x_n - x_{n+1}) + 2\lambda \sigma_n^2 x_n. \tag{4.7}$$

Let us compute the Hessian matrix of J_λ,

$$\frac{\partial^2}{\partial x_n \partial x_m} J_\lambda = \begin{cases} 2\eta\sigma_{n-1}/V_{n-1} + 2\eta\sigma_n/V_n + 2\lambda\sigma_n^2 & \text{if } m = n, \\ -2\eta\sigma_{n-1}/V_{n-1} & \text{if } m = n-1, \\ -2\eta\sigma_n/V_n & \text{if } m = n+1, \\ 0 & \text{otherwise.} \end{cases}$$

If $\eta = 0$ then the Hessian matrix of J_λ is diagonal and has positive eigenvalues, which implies that the unique critical point of J_λ is a global minimum. By continuity of the eigenvalues, if η is small then all critical points of J_λ are still minima.

To find the minimum of J_λ we equate (4.7) to zero. We thus obtain the optimal trading curve as a solution to the recursive algorithm

$$x_{n+1} = \left(1 + \frac{\mathcal{D}_{n-1}}{\mathcal{D}_n} + \lambda \frac{\sigma_n^2}{\mathcal{D}_n}\right) x_n - \frac{\mathcal{D}_{n-1}}{\mathcal{D}_n} x_{n-1}, \quad \mathcal{D}_n = \eta \frac{\sigma_n}{V_n}, \tag{4.8}$$

under the constraints $x_0 = 1$ and $x_{N+1} = 0$ (see Figure 1.3).

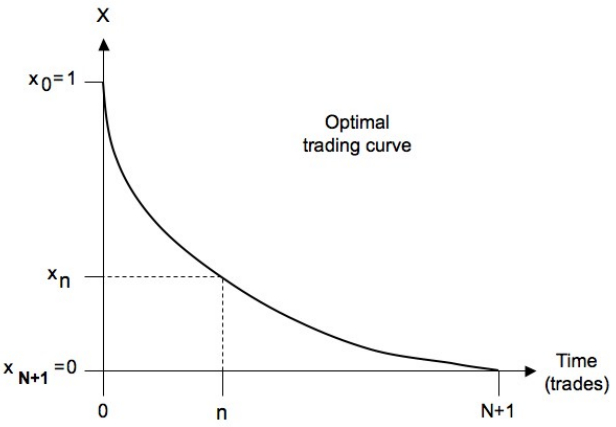

Figure 4.1: Optimal Trading curve. The curve describes the trading schedule that minimizes the joint effect of market impact and market risk.

4.2 Multi-asset portfolios

This section generalizes the previous ideas to the case of multi-asset Gaussian portfolios.

4.2.1 Model

We will consider here N trading dates and a portfolio of K assets $\mathbf{S} = (S^1, \dots, S^K)$. For any n we define the asset vector $\mathbf{S}_n = (S_n^1, \dots, S_n^K)$, whose dynamics is supposed to be Gaussian, i.e.

$$\mathbf{S}_{n+1} = \mathbf{S}_n + \mathcal{E}_{n+1}, \tag{4.9}$$

where $(\mathcal{E}_n)_{1 \le n \le N}$ are i.i.d. K-dimensional Gaussian vectors with mean zero and covariance matrix $\mathfrak{S}_n$. Following Almgren and Chriss [1] again, we will model the market impact as a function

$$h_n = h(\nu_n^k),$$

where

$$(\nu_1^k, \dots, \nu_N^k), \qquad \sum_{n=1}^{N} \nu_n^k = 1 \qquad \text{for all } k = 1, \dots, K$$

are the number of units of asset k that are traded at time n. We define as before x_n^k as the remaining quantity of asset k to buy at time n, i.e.

$$x_n^k = \sum_{i=n}^{N} \nu_n^k, \quad \nu_n^k = x_n^k - x_{n+1}^k.$$

Under this framework, we can calculate the wealth process (i.e. the full trading revenue upon completion of all trades):

$$W = \sum_{k=1}^{K} \sum_{n=1}^{N} \delta_n^k \nu_n (S_n^k + \delta_n^k h(\nu_n^k)). \tag{4.10}$$

As in the single-asset case we will suppose that $\delta_n^k = 1$ for all pairs (n, k) and that

$$h(\nu_n^k) = \eta^k \frac{\sigma_n^k \nu_n^k}{V_n^k}, \qquad \eta^k > 0, \tag{4.11}$$

with σ_n^k and V_n^k the volatility and the historical volume of the asset k at time n, respectively.

Repeating the arguments for the single-asset case it can be shown that the wealth process (4.12) in its simplified form is

$$W = \sum_{k=1}^{K} S_0^k + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathcal{E}_n^k x_n^k + \sum_{k=1}^{K}\sum_{n=1}^{N} \eta^k \frac{\sigma_n^k}{V_n^k}(x_n^k - x_{n+1}^k)^2. \tag{4.12}$$

4.2.2 Optimization program

Computing the expectation and the variance in (4.12) yields

$$\begin{aligned} \mathbb{E}(W) &= \sum_{k=1}^{K} S_0^k + \sum_{k=1}^{K}\sum_{n=1}^{N} \eta^k \frac{\sigma_n^k}{V_n^k}(x_n - x_{n+1})^2, \\ \mathbb{V}(W) &= \sum_{n=1}^{N} \mathbf{x}_n' \mathfrak{S}_n \mathbf{x}_n, \quad \mathbf{x}_n = (x_n^1, \ldots, x_n^K). \end{aligned} \tag{4.13}$$

We are looking for the strategy $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ that minimizes the cost functional

$$\begin{aligned} J_\lambda(\mathbf{x}_1, \ldots, \mathbf{x}_N) &= \mathbb{E}(W) + \lambda \mathbb{V}(W) \\ &= \sum_{k=1}^{K} S_0^k + \sum_{n=1}^{N} \left(\lambda \mathbf{x}_n' \mathfrak{S}_n \mathbf{x}_n + \sum_{k=1}^{K} \eta^k \frac{\sigma_n^k}{V_n^k}(x_n - x_{n+1})^2 \right). \end{aligned} \tag{4.14}$$

As in the single-asset case, if $\eta^k = 0$ for all $k = 1, \ldots, K$ then the unique critical point of J_λ is a minimum. In consequence, if the vector $\eta = (\eta^1, \ldots, \eta^K)$ is sufficiently small then the critical points of J_λ are still minima.

Notice that J_λ is a functional of the variables x_n^k. Therefore, in order to find its minimum we have to solve the $K \times N$ scalar equations

$$\frac{\partial J_\lambda}{\partial x_n^k} = 0, \qquad k = 1, \ldots, K \qquad n = 1, \ldots N.$$

This is equivalent to solving the N vectorial equations for the corresponding gradients, i.e.

$$\frac{\partial J_\lambda}{\partial \mathbf{x}_n} = 0, \qquad n = 1, \ldots N.$$

After some computations we obtain the recursive algorithm

$$\mathbf{x}_{n+1} = \left(1 + \mathcal{D}_n^{-1}\mathcal{D}_{n-1} + \lambda \mathcal{D}_n^{-1}\mathfrak{S}_n\right)\mathbf{x}_n - \mathcal{D}_n^{-1}\mathcal{D}_{n-1}\mathbf{x}_{n-1}\,, \tag{4.15}$$

where $\mathcal{D}_n$ is the $K \times K$ diagonal matrix with kth element

$$\mathcal{D}_n^k = \frac{\eta^k \sigma_n^k}{V_n^k}\,.$$

Therefore, the optimal trading curve solves the recursive algorithm (4.15) for $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ under the constraints $x_0^k = 1$ and $x_{N+1}^k = 0$ for all $k = 1, \ldots, K$.

4.3 Balanced portfolios

We will study now a particular case of multi-asset Gaussian portfolio, the so-called *balanced portfolio*. Let us suppose that that there is a unique trading strategy

$$(\tilde{\nu}_1, \ldots, \tilde{\nu}_N), \qquad \sum_{n=1}^{N} \tilde{\nu}_n = 1, \tag{4.16}$$

such that the trading schedule of any asset k is a constant multiple of it. More precisely, we will assume that for any $k = 1, \ldots, K$ there exists $\pi^k > 0$ such that

$$\nu_n^k = \pi^k \tilde{\nu}_n \qquad \text{for all } n = 1, \ldots, N.$$

π^k is thus the total number of shares of asset k to be traded. Under this framework we also have

$$\begin{aligned} x_n^k &= \sum_{i=n}^{N} \nu_i^k = \sum_{i=n}^{N} \pi^k \tilde{\nu}_i \\ &= \pi^k \sum_{i=n}^{N} \tilde{\nu}_i = \pi^k \tilde{x}_n. \end{aligned} \tag{4.17}$$

Define $\pi = (\pi^1, \ldots, \pi^K)$. From (4.16) and (4.17) it follows that the wealth process (4.12) for a balanced portfolio takes the form

$$W = \sum_{n=1}^{N} S_0^n + \sum_{n=1}^{N} \tilde{x}_n \left(\sum_{k=1}^{K} \mathcal{E}_n^k \pi^k \right) + \sum_{n=1}^{N} (\tilde{x}_n - \tilde{x}_{n+1})^2 \left(\sum_{k=1}^{K} \eta^k (\pi^k)^2 \frac{\sigma_n^k}{V_n^k} \right),$$

whereas the cost functional (4.14) becomes

$$J_\lambda(\tilde{x}_1, \ldots, \tilde{x}_N) = \sum_{k=1}^{K} S_0^k + \lambda \sum_{n=1}^{N} (\tilde{x}_n)^2 \left(\pi' \mathfrak{S}_n \pi \right) + \sum_{n=1}^{N} (\tilde{x}_n - \tilde{x}_{n+1})^2 \left(\sum_{k=1}^{K} \eta^k (\pi^k)^2 \frac{\sigma_n^k}{V_n^k} \right).$$

As we have already seen before, if $\eta = (\eta^1, \ldots, \eta^K)$ is small then the minimum of J_λ exists. After some calculations it can be shown that the optimal trading curve $(\tilde{x}_1, \ldots, \tilde{x}_N)$ solves the recursive algorithm

$$\tilde{x}_{n+1} = \left(1 + \frac{\mathcal{D}_{n-1}^\pi}{\mathcal{D}_n^\pi} + \lambda \frac{\pi' \mathfrak{S}_n \pi}{\mathcal{D}_n^\pi} \right) \tilde{x}_n - \frac{\mathcal{D}_{n-1}^\pi}{\mathcal{D}_n^\pi} \tilde{x}_{n-1}\,, \qquad \mathcal{D}_n^\pi = \sum_{k=1}^{K} (\pi^k)^2 \eta^k \frac{\sigma_n^k}{V_n^k}\,, \tag{4.18}$$

under the constraints $\tilde{x}_1 = 1$ and $\tilde{x}_{N+1} = 0$.

Chapter 5

Optimal trading for mean-reverting portfolios

In this chapter we will still assume that the assets still follow a Gaussian diffusion. However, we will suppose that there exists a non-gaussian asset such that the corresponding portfolio has a mean-reverting dynamics.

5.1 General mean-reverting portfolios

5.1.1 The model

We start with a multi-asset Gaussian portfolio $S = (S^1, \dots, S^K)$ with dynamics

$$S_{n+1} = S_n + \mathcal{E}_{n+1}, \qquad n = 1, \dots, N$$

where $(\mathcal{E}_n)_{1 \leq n \leq N}$ are i.i.d. K-dimensional Gaussian vectors with covariance matrix $\mathfrak{S}_n$. Suppose that we decide to trade portfolio S in balanced mode and that there exists an asset A (with finite standard deviation σ^A) such that the portfolio

$$C_n = \sum_{k=1}^{K} \pi^k S_n^k + A_n$$

has a mean-reverting dynamics. We will assume that C follows an Ornstein-Uhlenbeck process, i.e.

$$\mathrm{d}C_t = \gamma(M - C_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{5.1}$$

where M is the mean. The speed parameter γ determines how fast does the process converge to its mean. The discrete version of (5.3) is the so-called *auto-regressive process* of order one (AR(1)),

$$C_{n+1} = C_n + \gamma(M - C_n) + \sigma\varepsilon_{n+1}, \tag{5.2}$$

where $(\varepsilon_n)_{1 \leq n \leq N}$ are i.i.d normal random variables. For the sake of simplicity we will suppose that $M = 0$. Under this considerations, (5.2) becomes

$$C_{n+1} = (1 - \gamma)C_n + \sigma\varepsilon_{n+1}. \tag{5.3}$$

Finally, using induction on (5.3) yields

$$C_n = (1-\gamma)^n C_0 + \sum_{i=1}^{n} (1-\gamma)^{n-i} \sigma \varepsilon_i. \tag{5.4}$$

5.1.2 Wealth process and optimization program

We normalize portfolio C, i.e.

$$C_n = \sum_{k=1}^{K} \frac{\pi^k}{m} S_n^k + \frac{1}{m} A_n, \qquad m = \sum_{k=1}^{K} \pi^k + 1,$$

and assume that portfolio C is balanced, i.e. for all n we have

$$\nu_n^k = \frac{\pi^k \nu_n}{m} \quad k = 1, \ldots, K \qquad \text{and} \qquad \nu_n^A = \frac{\nu_n}{m}.$$

As in the previous chapter, we will suppose that the portfolio is long in all assets and that the market impact is

$$\begin{aligned} &\eta^k \frac{\sigma_n^k \nu_n^k}{V_n^k} && \text{for asset } S^k,\ k = 1, \ldots, K \\ &\eta^A \frac{\sigma_n^A \nu_n^A}{V_n^A} && \text{for asset } A. \end{aligned}$$

We readily compute,

$$\begin{aligned} \sum_{n=1}^{N} \nu_n C_n &= \sum_{n=1}^{N} \nu_n \left((1-\gamma)^n C_0 + \sum_{i=1}^{n} (1-\gamma)^{n-i} \sigma \varepsilon_i \right) \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} \sum_{i=1}^{n} \nu_n (1-\gamma)^{n-i} \sigma \varepsilon_i \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} \sum_{i=n}^{N} \nu_i (1-\gamma)^{i-n} \sigma \varepsilon_n \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} \sigma \varepsilon_n \left(\sum_{i=n}^{N} \nu_i (1-\gamma)^{i-n} \right) \end{aligned}$$

and we find that the wealth process W takes the form

$$W = \sum_{n=1}^{N} \frac{\nu_n}{m} (1-\gamma)^n C_0 + \sum_{n=1}^{N} \frac{\sigma \varepsilon_n}{m} \left(\sum_{i=n}^{N} \nu_i (1-\gamma)^{i-n} \right) + \sum_{n=1}^{N} b_n (\nu_n)^2 , \tag{5.5}$$

where

$$b_n = \frac{1}{m^2} \left(\sum_{k=1}^{K} \eta^k (\pi^k)^2 \frac{\sigma_n^k}{V_n^k} + \eta^A \frac{\sigma_n^A}{V_n^A} \right).$$

Unfortunately, the wealth process $W(\nu_1, \ldots, \nu_N)$ given in (5.5) does not admit a reduced representation in terms of $(x_1, \ldots, x_N)$, which implies that there is no explicit recursive algorithm for the optimal trading curve. In consequence, in order to construct the optimal trading curve it is necessary find a minimum of the cost functional $J_\lambda(\nu_1, \ldots, \nu_N)$ via an optimization program in $\mathbb{R}^N$ under the constraints

$$\nu_n \geq 0 \quad \text{for all } n = 1, \ldots, N \quad \text{and} \quad \sum_{n=1}^{N} \nu_n = 1.$$

5.2 Simplified model

We would like to have some insight on the shape of the optimal trading curve for mean-reverting portfolio. For that purpose, let us consider a simplified mean-reverting model.

Recall the recursive form for C_n given in (5.4) and observe that it can also be expressed as

$$C_n = (1-\gamma)^n C_0 + \sum_{i=1}^{n-1} (1-\gamma)^{i-1} \sigma \varepsilon_{n-i+1} \,. \tag{5.6}$$

We propose a new portfolio process C_n satisfying

$$\tilde{C}_n = (1-\gamma)^n C_0 + \sum_{i=1}^{n} (1-\gamma)^{i-1} \sigma \varepsilon_i \,. \tag{5.7}$$

Notice that our simplified model (5.7) is just the original AR(1) model given in (5.6) after a permutation on the normal random variables, i.e.

$$\begin{aligned} \tilde{C}_n &\leftrightarrow C_n \\ \varepsilon_i &\leftrightarrow \varepsilon_{n-i+1} \,. \end{aligned}$$

Unlike the full mean-reverting model, the new process (5.7) does admit a wealth process W with explicit dependence on $(x_1, \ldots, x_N)$. Indeed,

$$\begin{aligned} \sum_{n=1}^{N} \nu_n \tilde{C}_n &= \sum_{n=1}^{N} \nu_n \left((1-\gamma)^n C_0 + \sum_{i=1}^{n-1} (1-\gamma)^{i-1} \sigma \varepsilon_i \right) \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} \sum_{i=1}^{n} \nu_n (1-\gamma)^{i-1} \sigma \varepsilon_i \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} \sum_{i=n}^{N} \nu_i (1-\gamma)^{n-1} \sigma \varepsilon_n \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} (1-\gamma)^{n-1} \sigma \varepsilon_n \left(\sum_{i=n}^{N} \nu_i \right) \\ &= \sum_{n=1}^{N} \nu_n (1-\gamma)^n C_0 + \sum_{n=1}^{N} (1-\gamma)^{n-1} \sigma \varepsilon_n x_n \,. \end{aligned}$$

We thus obtain that the wealth process is

$$W = \sum_{n=1}^{N} \frac{(x_n - x_{n+1})}{m}(1-\gamma)^n C_0 + \sum_{n=1}^{N} \frac{x_n}{m}(1-\gamma)^{n-1}\sigma_n \varepsilon_n + \sum_{n=1}^{N} b_n (x_n - x_{n+1})^2 \,, \tag{5.8}$$

where

$$b_n = \frac{1}{m^2}\left(\sum_{k=1}^{K} \eta^k (\pi^k)^2 \frac{\sigma_n^k}{V_n^k} + \eta^A \frac{\sigma_n^A}{V_n^A}\right).$$

Computing the expectation and the variance of the wealth function W we obtain the cost functional

$$J_\lambda(x_1, \ldots, x_N) = \sum_{n=1}^{N} \frac{C_0}{m}(1-\gamma)^n (x_n - x_{n+1}) + \lambda \sum_{n=1}^{N} \frac{\sigma^2}{m^2}(1-\gamma)^{2n-2}(x_n)^2 + \sum_{n=1}^{N} b_n (x_n - x_{n+1})^2. \tag{5.9}$$

As we have already seen before, if $\eta = (\eta^1, \ldots, \eta^K, \eta^A)$ is small enough then J_λ has a minimum. In order to find this minimum, we compute the partial derivatives of J_λ,

$$\frac{\partial J_\lambda}{\partial x_n} = -\gamma(1-\gamma)^{n-1} C_0 + \frac{2\lambda}{m^2}(1-\gamma)^{2n-2}\sigma^2 x_n + 2b_n(x_n - x_{n+1}) - 2b_{n-1}(x_{n-1} - x_n)\,,$$

and equate them to zero. We thus obtain the following recursive algorithm for the optimal trading curve,

$$x_{n+1} = \gamma(1-\gamma)^{n-1} U_n + \left(1 + \lambda(1-\gamma)^{2n-2} Z_n + B_n\right) x_n - B_n x_{n-1}, \tag{5.10}$$

where

$$U_n = -\frac{C_0}{2b_n m}, \qquad Z_n = \frac{\sigma^2}{b_n m^2}, \qquad B_n = \frac{b_{n-1}}{b_n},$$

which has to be solved under the constraints $x_0 = 1$, $x_{N+1} = 0$.

In Section 5.3 we will make a detour to the realm of differential equations in order to describe a numerical method for solving recursive algorithms, called the *shooting method*. In Section 5.4 we will program the shooting method in Matlab and construct the optimal trading curve associated to the algorithm (5.10).

5.3 The shooting method

5.3.1 Description

Consider the initial-value problem

$$y'' = g(y, y'); \quad t \in [a, b], \quad y(a) = A, \quad y'(a) = \alpha, \tag{5.11}$$

where g is a bounded and differentiable function. According to the standard theory of Ordinary Differential Equations (ODE), the initial value problem (5.11) has a unique solution $y(t)$.[1]

Now consider the boundary problem

$$y'' = g(y, y'); \quad t \in [a, b], \quad y(a) = A \in \mathbb{R}, \quad y(b) = B \in \mathbb{R}. \tag{5.12}$$

It is not evident that (5.12) has a solution. However, we can try to translate the boundary problem (5.12) into an initial-value problem of type (5.11), for which we know that solutions do exist.

The shooting method consists exactly in this translation. Given $\alpha \in \mathbb{R}$, the initial-value problem (5.11) has a solution $y(t, \alpha)$. To solve the boundary problem (5.12), we need to find α_0 such that $y(b; \alpha_0) = B$. Roughly speaking, we are playing with the" shooting angle" α in order to "hit" B at $t = b$ (see Figure 5.1).

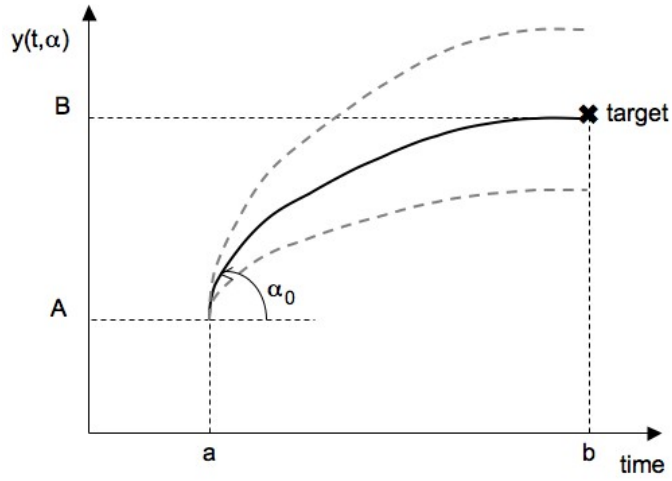

Figure 5.1: The shooting method. Varying the "angle of shooting" α we can find the right angle α_0 such that the curve $y(t, \alpha)$ that "hits the target", i.e. $y(b; \alpha_0) = B$.

In consequence, the boundary problem (5.12) reduces to find a zero of the function

$$F(\alpha) = y(b, \alpha) - B,$$

which can be solved using any numerical method, e.g. bisection or Newton (see Stoer and Bulirsch [22] for mode details).

The shooting method is also valid for systems of second-order ODEs. Indeed, the initial-value problem

$$\mathbf{y}'' = g(\mathbf{y}, \mathbf{y}'); \quad t \in [a, b], \quad \mathbf{y}(\mathbf{a}) = \mathbf{A} \in \mathbb{R}^K, \quad \mathbf{y}'(\mathbf{a}) = \alpha \in \mathbb{R}^K,$$

[1] Strictly speaking, the solution $y(t)$ only exists locally, but it is globally defined if g and all its partial derivatives are continuous and bounded. See Perko [20] for more details.

admits a unique solution $\mathbf{y}(t, \alpha)$, which permits to define the functional

$$F(\alpha) = \mathbf{y}(\mathbf{b}, \alpha) - \mathbf{B}.$$

5.3.2 Application to optimal trading curves

Suppose we have already found the optimal trading curve $(x_1, \ldots, x_N)$ via a recursive algorithm of the form, i.e.

$$x_{n+1} = F(x_n, x_{n-1}) \tag{5.13}$$

under the constraints $x_0 = 1$ and $x_{N+1} = 0$. Using an induction argument it can be shown that x_n is a function of x_0 and x_1 for $n \geq 2$. By induction we obtain that x_{N+1} is a function of x_1, i.e.

$$x_{N+1} = F(x_1),$$

because x_0 has been already fixed to be equal to 1. If we define $x_1 = \alpha$ then α is a free parameter that completely determines the optimal trading curve. An important remark is that α can be related to the the slope of the trading curve at x_0 because

$$\frac{x_1 - x_0}{1 - 0} = \alpha - 1.$$

By considering α as the "slope" we can see an analogy of the optimal trading curve with the shooting method. Under this new framework, our optimization problem reduces to find a zero of $x_{N+1} = F(\alpha)$, i.e. a number α_0 such that $F(\alpha_0) = 0$.

The beauty of the analogy with the shooting method is that we are working with a 1-dimensional function $F(x_1)$ instead of the N-dimensional functional $J_\lambda(x_1, \ldots, x_N)$. In consequence, using the shooting method we are always solving a 1D problem regardless of the number of trades N. This fact renders our algorithm very appealing for high frequency trading.

In summary, the shooting method can be used to find optimal trading curves as long as the optimization program admits a recursive algorithm of the form

$$\mathbf{x}_{n+1} = F(\mathbf{x}_n, \mathbf{x}_{n-1}),$$

where $\mathbf{x}_0 \in \mathbb{R}^K$ is given and $\mathbf{x}_{N+1} = (0, \ldots, 0)$. This condition is rather general because it is satisfied by a large class of algorithms, e.g. single asset (4.8), balanced portfolio, multi-asset portfolio (4.15), (4.18) and simplified mean-reverting (5.10).

5.4 Numerical example using Matlab

We solved numerically the trading algorithm (5.10),

$$x_{n+1} = \gamma(1-\gamma)^{n-1}U + (2 + \lambda(1-\gamma)^{2n-2})x_n - x_{n-1},$$

with the constraints $x_0 = 1$ and $x_{N+1} = 0$. We programmed the shooting method in Matlab for the values $U = 0.1$, $\lambda = 0.2$ and $N = 100$. The free parameter $\alpha = x_1$ lies in $[0, 1]$. Since for any

$\gamma \in (0,1)$ we found that $F(0) < 0$ and $F(1) > 0$, we only need to construct a simple bisection method (i.e. nested intervals) in order to find the zero of the function F.

Let us describe the bisection method we used:

- We start with the interval $I_1 = [0,1]$ with $F(0) < 0$ and $F(1) > 0$.
- If $F(1/2) > 0$ then we choose the interval $I_2 = [0, 1/2]$ because F changes sign inside. If $F(1/2) < 0$ then the sign changes in $I_2 = [1/2, 1]$
- Given the interval $I_j = [a_j, b_j]$ with $F(a_j) < 0$ and $F(b_j) > 0$, define $r = (a_j + b_j)/2$, i.e. the mid-point of I_j. If $F(r) > 0$ then define $I_{j+1} = [a_j, r]$, whereas if $F(r) < 0$ then $I_{j+1} = [r, b_j]$.
- Proceeding this way we can find α_0 such that $F(\alpha_0) = 0$.
- The optimal trading curve is now completely determined using $x_0 = 1$ and $x_1 = \alpha_0$. Notice that by construction we necessarily have $x_{N+1} = 0$.

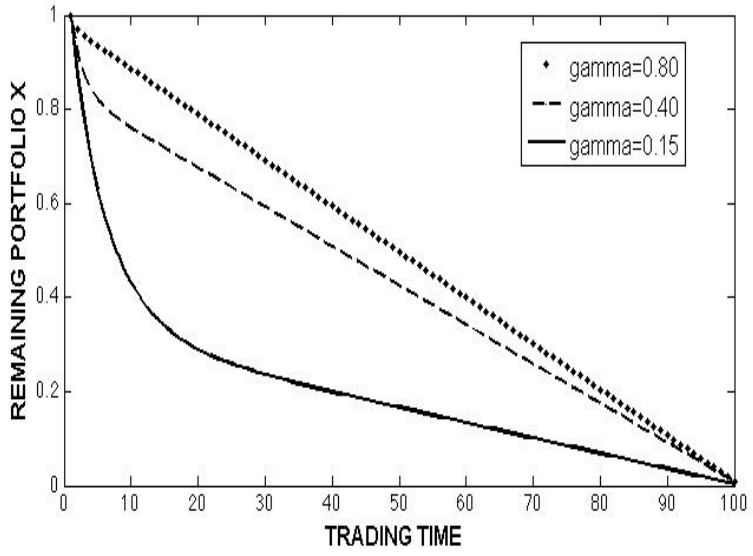

Figure 5.2: Optimal trading curves as a function of the speed parameter γ.

The graph of the optimal trading curve for three different values of γ is given in Figure 5.2. From the shape of the curves we can infer the following facts:

- For big γ (e.g. $\gamma = 0.80$) the mean-reverting process is stronger than the diffusion process. Therefore, we are only minimizing the market impact, which implies that the trading strategy looks like the a straight line.

- However, for small γ (e.g. $\gamma = 0.15$) the mean-reverting process is almost inexistent, i.e. we have only a diffusion process. Hence, the optimal strategy is to trade more at the beginning of the execution in order to minimize the market risk.

In the graph of the trading curve with $\gamma = 0.15$ we can see that there are two distinct trading patterns:

- At the beginning of the execution we minimize the market risk by trading as fast as possible. At $n = 15$ we have already executed 70% of the order.
- For the rest of the execution (i.e. the remaining 85 trades) we minimize the market impact by trading the remaining 30% of the portfolio as slow as possible.

When we switch from a trading pattern to the other, the amount of shares x remaining in the portfolio is an increasing function of γ: if γ is small then x is small, and vice-versa. This two-fold pattern is also visible for $\gamma = 0.40$, but when $\gamma \to 1$ the market risk disappears and we have only the trading pattern that minimizes market impact. Such a behavior was to be expected from a dynamics similar to an Ornstein-Uhlenbeck process because for short times the leading term is of order $\sqrt{t}$, i.e. the diffusion term $\sigma \mathrm{d}W_t$, whereas for large times the leading term is of order t, i.e. the drift term $\gamma(M - C_t)\mathrm{d}t$, which in the long run converges to the mean M.

Chapter 6

Conclusions

6.1 Optimal trading curves

6.1.1 Recursive algorithms and shooting method

As we have shown, the theoretical framework for optimal trading curves is analogue to the efficient frontier, which is one of core results of Modern Portfolio Theory (MPT) and Capital Asset Pricing Model (CAPM). The idea of the efficient frontier comes from the fact that investors want to maximize his return given a risk budget, or equivalently to minimize his risk given a target or fixed return.

In this work we explicitly computed the optimal trading curves for single assets and portfolios. Moreover, they are algebraic and have a nice graphic interpretation in terms of the trading schedule of the portfolio, which grants flexibility to the trader and lets him monitor the execution of his trading schedule at all times, so he can adjust the parameters on the run if necessary. The same approach holds for other trading strategies, and could also be applied for other market models.

We have used the shooting method, a well-known numerical technique in differential equations, to solve the optimization of the trading curve. With this technique we can find optimal trading curve as long as the optimization program admits an explicit recursive algorithm of the form

$$\mathbf{x}_{n+1} = F(\mathbf{x}_n, \mathbf{x}_{n-1}),$$

where $\mathbf{x}_0$ is given and $\mathbf{x}_{N+1} = (0, \ldots, 0)$.

6.1.2 Dynamic programming and optimal control

If a recursive algorithm is not possible, we have to try a different approach. Our first suggestion is the method of dynamic programming.

Several problems in Finance and Economics involve the maximization of a utility function (or the minimization of a cost/hedging function) under some constraints. When the functions are

deterministic the optimization program is easily achieved using Lagrange multipliers as before; however, if the utility function depends explicitly on stochastic processes we need to maximize it over a certain range of possible choices (called *controls*).

In the Black-Scholes framework, there is a unique hedging strategy, which leads to a linear partial differential equation (PDE). In practice, however, there are cases where the hedging strategy is not unique and other where it does not even exist. In those non-Black-Scholes scenarios we end up dealing with more complicated PDEs, either nonlinear equations or inequalities, which solutions (called viscosity solutions) can only be found numerically (see Evans [8]).

The application of the dynamic programming principle in Bouchard *et al* [4] involves stochastic targeting, where the admissible controls satisfy some given constraints (e.g. a fixed number of trades, a maximum/minimum level of volume per trade and/or fixed trading dates). This leads to complicated but fascinating mathematics, which is currently a fertile ground of research.

6.1.3 Nonlinear transaction costs

In the seminal work of Almgren and Chriss [1], the liquidity cost per share traded is a linear function of trading rate or of block size, and that the only source of uncertainty in execution is the volatility of the underlying asset. Researchers are currently looking for more complex transaction cost functions. Almgren [2] proposed nonlinear transaction cost functions, where market impact cost per share are a power law function of the trading rate. Almgren *et al* [3] analyzed real data and found that market impact is a power law of exponent 3/5 of block size, with specific dependence on trade duration, daily volume and volatility.

For the reader interested in these nonlinear models for transaction costs we suggest to start with Bouchaud and Potters [7], Almgren *et al* [3] and the references therein.

6.2 Normal returns vs real returns: stylized facts

According to the Efficient Market Theory (EMT), the future states of world markets and their respective probabilities are known. In other words, the distribution of prices at any time is known by all and every agent. However, instead of estimating the distribution directly from the empirical data, the mainstream economic theory supposes *a priori* that the distribution is (log)normal. This is scientifically wrong because it is the facts that determine the models, not vice versa.

After studying carefully the real distributions of financial returns, the empirical evidence shows that, despite the heterogeneity of the assets (commodities, fixed income, foreign exchange, equities, credit), they all share the same properties. These pattern are very recurrent, almost ubiquitous, and they appear regardless of the time-scale of the returns (intra-day, week, month, year). For all these reasons, the distinctive characteristics of real returns have been elevated to the status of facts: they are called *stylized facts* (Embrechts *et al* [19]).

The real distribution of any financial assets has the following stylized facts, as it is shown in Embrechts *et al* [19] and Veredas [23] :

- It is approximately symmetric.
- It has fat tails, i.e. extreme returns are more likely to happen that the normal theory would forecast.
- It has a high peak (i.e. *leptokurtic*).
- There is weak autocorrelation between the returns for different days, i.e. conditional expected returns are close to zero.
- There is a high autocorrelation in the fluctuations of returns, i.e. in the absolute value of square returns.
- Volatility varies over time and presents clustering patterns, i.e. there are periods of high volatility and others of low volatility.

6.3 Some alternative models in Economics and Finance

It is important to remark that a normal distribution does not satisfy any of these properties. Therefore, it is necessary to use more accurate distributions to approach real financial markets. There are several models that are becoming serious alternatives to the (flawed) normal distribution.

6.3.1 GARCH

We want to to take into account the stylized facts, but perhaps we do not want to abandon completely the idea of normal returns. In that case we could suppose a normal distribution, not on the unconditional returns but on the conditional ones, and that the volatility at a given time depends on past volatilities and returns. A natural step forward is the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model, published by Robert F. Engle in 1982.[1]

Suppose that today is time t and that the information available is up to yesterday, i.e. $\mathcal{F}_{t-1}$. We will assume that the conditional returns today $r_t|\mathcal{F}_{t-1}$ are normal, and that today's volatility t depends on the volatility and return levels of yesterday, i.e.

$$\begin{aligned} r_t|\mathcal{F}_{t-1} &\sim N(\mu, h_t), \\ h_t|\mathcal{F}_{t-1} &= h(r_{t-1}, h_{t-1}). \end{aligned}$$

More precisely, the conditional returns and volatility have the form

$$\begin{aligned} r_t &= \mu + h_{t-1}^{1/2} z_t, \quad z_t \sim N(0,1) \quad \text{i.i.d.} \\ h_t &= \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1}\,. \end{aligned} \tag{6.1}$$

[1] Robert F. Engle won the Nobel prize in Economics in 2003 for his GARCH model.

Model (6.1) satisfies all the stylized facts : returns are not autocorrelated yet they are not independent, there is volatility clustering and the unconditional GARCH distribution has more kurtosis than the normal distribution.

There are several generalizations of the GARCH model :

- Asymmetric versions of GARCH : they have non-zero skewness.
- GARCH(p, q) : The conditional volatility depends on several past levels of volatility and returns :

$$\begin{aligned} r_t &= \mu + h_{t-1}^{1/2} z_t, \quad z_t \sim N(0,1) \quad \text{i.i.d.} \\ h_t &= \omega + \sum_{i=1}^{q} \alpha_i (r_{t-i} - \mu)^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \,. \end{aligned} \tag{6.2}$$

6.3.2 Lévy distributions

The tails of a Gaussian distribution decay exponentially fast :

$$N(x) \underset{x \to \pm\infty}{\sim} \exp\left(-x^2\right).$$

Therefore, any distribution whose tails decay slower is a good candidate for a fat tail modeling. In that sense, Lévy (also called Pareto) distributions are a natural choice because they decay following a power-law (see e.g. Bouchaud and Potters [7]) :

$$L_\mu(x) \underset{x \to \pm\infty}{\sim} \frac{\mu A_\pm^\mu}{|x|^{1+\mu}}, \quad 0 < \mu < 2, \quad A_\pm^\mu > 0.$$

Some Lévy distributions carry an asymmetry parameter β that measures the relative weight of the positive and negative tails. Since in the limit $\mu = 2$ we recover the Gaussian distribution, Lévy distributions are generalizations of normal distributions. However, their tails are a bit too fat: for $\mu < 2$ the variance is infinite and for $\mu \leq 1$ even the mean is infinite. In order to overcome this problem we can use a *truncated Lévy distribution*, which is Lévy in an intermediate regime $|x| << +\infty$ and exponential at the tails (hence still fatter than the Gaussian) :

$$P(x) \underset{x \to \pm\infty}{\sim} \exp(-\alpha|x|), \quad \alpha > 0$$

This allows to have finite moments of all order, and in particular the most useful ones in Finance : first (the mean), second (variance), third (skewness) and fourth (kurtosis).

6.3.3 Student distributions

Besides truncated Lévy distributions, Student distributions are also compatible with the stylized facts,

$$S_\nu(x) = c(\nu) \left[1 + \frac{x^2}{\nu}\right]^{-(1+\nu)/2}, \quad 1 \leq \nu < +\infty,$$

where $c(\nu)$ is a normalizing constant. A remarkable property of the Student distribution is that the parameter ν determines the tail behavior : the tails got heavier as ν decreases. Since in the limit $\nu \to +\infty$ the distribution $S_\mu(x)$ tends towards the normal, Student distributions are generalizations of Gaussian distributions. Moreover, unlike the pure Lévy distributions, the parameter ν can be chosen such that the real distribution and the Student one have exactly the same finite (resp. inifinte) moments (see Bouchard and Potters [7]). Using the Student distribution we can construct a non-Gaussian GARCH model, where the residuals are no longer Gaussian but Student :

$$\begin{aligned} r_t &= \mu + h_{t-1}^{1/2} z_t, \quad z_t \sim \text{Student}(0,1,\nu) \text{ i.i.d.} \\ h_t &= \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1}\,. \end{aligned} \tag{6.3}$$

6.3.4 Fractional Brownian motion

One of the properties of the Brownian motion B (i.e. normal increments) is that they are proportional to the square root of the time step, i.e.

$$B(t) - B(s) \propto |t-s|^{1/2}.$$

The fractional Brownian motion B^H of Hurst exponent $H \in (0,1)$ is a generalization of this fact :

$$B^H(t) - B^H(s) \propto |t-s|^H.$$

The process B^H possess the following properties (Bouchaud and Potters [7] and Mandelbrot [18]) :

- If $H = 1/2$ the returns are not correlated and B^H is the Brownian motion (also called *white noise*).
- If $H > 1/2$ the returns are positively correlated (and the process is called *red noise*).
- If $H < 1/2$ the returns are negatively correlated (*blue noise*).

Mandelbrot [18] reckons that there are a lot of financial assets whose returns are positively correlated (e.g. Apple). This fact is a flagrant violation of the EMT, which cannot explain the phenomenon. Other phenomena outside the scope of the EMT (yet always present in financial markets) are bubbles, crashes, herd behavior and long-run trends.

6.3.5 Multifractal Models

According to Mandelbrot [18], trading time is relative: there are periods of high market activity when news do not stop arriving and prices change very rapidly, but there are also periods of very few market activity when prices are rather quiet and news are scarce (see Mandelbrot [18]). A multifractal model can be constructed as a mix of two fractals : a fractal *mother* determining the dynamics of market prices and a fractal *father* representing the trading time relativity. Mandelbrot's choice for the fractal mother and father is the fractional Brownian Motion, whose Hurst exponents can be calibrated using market data. The main feature of multifractals is

that, unlike fractional Brownian motion and classical fractals, they are not scale-invariant by construction. In consequence, short-term and long-term behaviors are different. This is crucial for understanding the dynamics of real markets and microstructure effects in trading patterns.

6.3.6 Adaptive markets and agent-based models

A very clever way to think about Economics in general and financial markets in particular is borrowed from the Theory of Eolution. Andrew Lo and Doyne Farmer [9], [17] consider that we should be looking at financial markets from a biological perspective, i.e. within an evolutionary point of view. In their opinion, markets, instruments and participants interact and evolve dynamically according to the "law" of Economic Selection (i.e. the equivalent to Natural Selection). Financial agents compete, learn and adapt, but not necessarily in an optimal fashion (there is asymmetry of information and emotions that bias the decision-making process of each individual). When market condition change, new trading strategies emerge and replace the old ones that are obsolete for the current context. Eventually, the new strategies will in turn become obsolete and be replaced; and the evolutionary cycle will continue *ad infinitum*.

6.4 Taking a stand: quantitative vs discretionary trading

There is an interesting debate between quantitative and discretionary traders. There are several seminars and discussion groups on the topic in universities, financial firms and the Internet. The idea of technology eventually replacing humans is very recurrent in every field. In trading, however, it is not far from reality because trading floors and brokers are becoming electronic and automated.

I would like to draw the reader's attention to the online announcement for a seminar on quantitative trading that will be held next autumn in Paris, France (*Quant Invest 2010*, November 29 - December 10) :

> "Manager vs Machine :
>
> The investment universe is densely populated with skilled managers all hunting for elusive alpha. Amongst these managers a new breed is emerging: Generation Q - the investment managers that are backed by an arsenal of quantitative techniques and strategies that consistently give them the edge over others in the market.
>
> Quantitative managers take a scientific approach to financial markets, stripping human emotions such as fear and greed out of the equation. Instinct and intuition play little part in strategies that employ immense computing power.
>
> The disciplined process that underpins quantitative investment strategies means that while the models themselves are complex the investment philosophy of a quantitative manager is highly transparent."

As long as the market in question is electronic and liquid, quantitative approaches are better positioned to capture short-living profit opportunities. Indeed, some of the most sophisticated

algorithms can monitor several hundreds of variables simultaneously, hence they beat any human trader in terms of reactivity and accuracy. However, there is a strong point supporting human trading. When markets are illiquid and/or OTC (as it is the case for credit derivatives such as CDS), human traders are necessary because prices and volumes are negotiated directly with dealers and market makers. Moreover, a trader has to anticipate the impact of news on the course of equities and credit spreads, which is not easy to quantify. Nevertheless, in order to build up a profitable trading book, any trader has to rely on some quantitative signals, models and forecasts of volatility, correlation, volume, market impact, transaction costs, spreads, etc. Indeed, there are a lot of players in the marked nowadays, which means that simple rules of thumb and easy profitable opportunities do not longer exist. In addition, the current trend in Finance favors the homogenization of OTC markets via clearing houses and standard products. This could eventually lead to more liquid credit markets, thus opening the gate for algorithmic trading.

In my opinion, all market participants will eventually become quantitative to a higher or lesser extent. The only ones that can be spared from the trend are long-term investors, but they will put their money in the hands of managers relying on quantitative signals and/or algorithms.

6.5 A final thought: how would be the trader of the future?

My personal opinion is that algorithms will substitute humans in the execution processes. Brokers are becoming more electronic and automated, and the trend will continue to widespread in the whole financial industry. My bet is that the "trader of the future" will be less operational and more intellectual: he will be less involved in eye-blinking reactivity of buy-sell gunfire type and much more involved in thinking, e.g. calibrating his algorithms *on the run* and creating new automated strategies. Therefore, human traders are not disappearing, they will just evolve and become more tech-friendly.

Let me try to explain better what I mean with"on the run". I am thinking on a semi-automated kind of trading, as in the following examples :

- An algorithm that can be easily adapted, e.g. changing the knock-in or knock-out thresholds at any time.
- An algorithm that you can switch off in order to trade by yourself for a while because there is new information coming and the AI will not be able to incorporate. After the opportunity is gone you turn the switch the algorithm on again.

It is like an airplane: when there is turbulence the human pilot takes full control, but in general he only monitors the automatic pilot. In the short run we will only have the sophisticated version of old trading style, but eventually the "rules of thumb" will be less and less useful due to automated trading, which is faster and more accurate when dealing with empirical rules. Therefore, I believe, in the long run humans will not be involved in trade-by-trade decision making, except in rare and profitable occasions when there is a need of inspiration, and much more involved in programming and calibrating algorithms.

Bibliography

[1] Robert Almgren, Neil Chriss, *Optimal execution of portfolio transactions.* Journal of Risk, Vol. 3, No. 2, 2010, pp. 5-39.

[2] Robert Almgren, *Optimal execution with nonlinear impact functions and trading-enhanced risk.* Applied Mathematical Finance 10, pp. 1-18.

[3] Robert Almgren, Chee Thum, Emmanuel Hauptman, Hong Li, *Equity market impact.* Risk, July 2005, pp. 57-62.

[4] Bruno Bouchard, Ngoc-Minh Dang and Charles-Albert Lehalle, *Optimal control of trading algorithms: a general impulse control approach.* Preprint.

[5] Jean-Philippe Bouchaud, *Economics needs a scientific revolution.* Nature, Vol. 45, p. 30, October 2008.

[6] Jean-Philippe Bouchaud, *The (unfortunate) complexity of the economy.* Physics World, April 2009.

[7] Jean-Philippe Bouchaud, Marc Potters, *Theory of financial risk and derivative pricing,* 2nd. edition, Cambridge, 2003.

[8] Laurence C. Evans, *An introduction to mathematical optimal control theory.* Course notes, University of California, Berkeley, 2010.

[9] J. Doyne Farmer, Andrew W. Lo, *Frontiers of finance: Evolution and efficient markets.* Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 9991-9992, August 1999.

[10] Carole Gresse, *Evaluation d'actifs.* Course notes, University Paris-Dauphine, 2010.

[11] Larry Harris, *Trading and exchanges : market microstructure for practitionners.* Oxford University Press, 2002.

[12] Terrence Hendershott, Ryan Riordan, *Algorithmic Trading and Information.* Working paper SSRN, September 2009.

[13] Barry Johnson, *Algorithmic trading and DMA.* 4Myeloma Press, USA, 2010.

[14] Charles-Albert Lehalle, *Rigorous strategic trading : balanced portfolio and mean reversion.* Journal of Trading, Summer 2009, pp. 40-46.

[15] Charles-Albert Lehalle, Mathieu Rosenbaum, *Trading haute fréquence.* Course notes, ENSAE, 2010.

[16] François-Serge Lhabitant, *Handbook of hedge funds.* John Wiley and Sons, 2006.

[17] Andrew W. Lo, *The Adaptive Market Hypothesis.* The Journal of Portfolio Management, 30th anniversary issue, pp. 15-29, 2004.

[18] Benoît Mandelbrot, *The (mis)behaviour of markets.* Profile Books, 2008.

[19] Alexander J. McNeil, Rüdiger Frey, Paul Embrechts, *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press, 2005.

[20] Laurence Perko, *Differential equations and dynamical systems.* Springer, USA, 2001.

[21] Fabrice Riva, *Microstructure des marchés financiers.* Course notes, University Paris-Dauphine, 2010.

[22] Josef Stoer, Roland Bulirsch, *Introduction to numerical analysis.* Springer, USA, 1983.

[23] David Veredas, *Financial time series.* Course notes, University Paris-Dauphine, 2010.